PHILIPPE RIGOLLET: OK, so the course you're currently sitting in is 18.650. And it's called Fundamentals of Statistics. And until last spring, it was still called Statistics for Applications. It turned out that really, based on the content, "Fundamentals of Statistics" was a more appropriate title.

I'll tell you a little bit about what we're going to be covering in class, what this class is about, what it's not about. I realize there's several offerings in statistics on campus. So I want to make sure that you've chosen the right one. And I also understand that for some of you, it's a matter of scheduling.

I need to actually throw out a disclaimer. I tend to speak too fast. I'm aware that.

Someone in the back, just do like that when you have no idea what I'm saying. Hopefully, I will repeat myself many times. So if you average over time, you'll see that statistics will tell you that you will get the right message that I was actually trying to stick to send.

All right, so what are the goals of this class? The first one is basically to give you an introduction. No one here is expected to have seen statistics before, but as you will see, you are expected to have seen probability. And usually, you do see some statistics in a probability course. So I'm sure some of you have some ideas, but I won't expect anything.

And we'll be using mathematics. Math class, so there's going to be a bunch of equations-- not so much real data and statistical thinking. We're going to try to provide theoretical guarantees. We have two estimators that are available for me-- how theory guides me to choose between the best of them, how certain can I be of my guarantees or prediction?

It's one thing to just bid out a number. It's another thing to put some error bars around. And we'll see how to build error bars, for example.

You will have your own applications. I'm happy to answer questions about specific applications. But rather than trying to tailor applications to an entire institute, I think we're going to work with pretty standard applications, mostly not very serious ones. And hopefully, you'll be able to take the main principles back with you and apply them to your particular problem.

What I'm hoping that you will get out of this class is that when you have a real-life situation-- and by "real life", I mean mostly at MIT, so some people probably would not call that real life-- their goal is to formulate a statistical problem in mathematical terms. If I want to say, is a drug effective, that's not in mathematical terms, I have to find out which measure I want to have to call it effective. Maybe it's over a certain period of time.

So there's a lot of things that you actually need. And I'm not really going to tell you how to go from the application to the point you need to be. But I will certainly describe to you what point you need to be at if you want to start applying statistical methodology. Then once you understand what kind of question you want to answer-- do I want a yes/no answer, do I want a number, do I want error bars, do I want to make predictions five years into future, do I have side information, or do I not have side information, all those things-- based on that, hopefully, you will have a catalog of statistical methods that you're going to be able to use and apply it in the wild.

And also, no statistical method is perfect. Some of the math people have agreed upon over the years, and people understand that this is the standard. But I want you to be able to understand what the limitations are, and when you make conclusions based on data, that those conclusions might be erroneous, for example.

All right, more practically, my goal here is to have you ready. So who has taken, for example, a machine-learning class here? All right, so many of you, actually-- maybe a third have taken a machine-learning class.

So statistics has somewhat evolved into machine learning in recent years. And my goal is to take you there. So machine learning has a strong algorithmic component.

So maybe some of you have taken a machine-learning class that displays mostly the algorithmic component. But there's also a statistical component. The machine learns from data.

So this is a statistical track. And there are some statistical machine-learning classes that you can take here. They're offered at the graduate level, I believe. But I want you to be ready to be able to take those classes, having the statistical fundamentals to understand what you're doing. And then you're going to be able to expand to broader and more sophisticated methods.

Lectures are here from 11:00 to 12:30 on Tuesday and Thursday. Victor-Emmanuel will also be-- and you can call him Victor-- will also be holding mandatory recitation. So please go on Stellar and pick your recitation. It's either 3:00 to 4:00 or 4:00 to 5:00 on Wednesdays. And it's going to be mostly focused on problem-solving.

They're mandatory in the sense that we're allowed to do this, but they're not going to cover entirely new material. But they might cover some techniques that might save you some time when it comes to the exam. So you might get by.

Attendance is not going to be taken or anything like this. But I highly recommend that you go, because, well, they're mandatory. So you cannot really complain that something was taught only in recitation. So please register on Stellar for which of the two recitations you would like to be in. They're capped at 40, so first come, first served.

Homework will be due weekly. There's a total of 11 problem sets. I realize this is a lot. Hopefully, we'll keep them light. I just want you to not rush too much.

The 10 best will be kept, and this will count for a total of 30% of the final grade. There are due Mondays at 8:00 PM on Stellar. And this is a new thing.

We're not going to use the boxes outside of the math department. We're going to use only PDF files. Well, you're always welcome to type them and practice your LaTeX or Word typing.

I also understand that this can be a bit of a strain, so just write them down on a piece of paper, use your iPhone, and take a picture of it. Dropbox has a nice, new-- so try to find something that puts a lot of contrast, especially if you use pencil, because we're going to check if they're readable. And this is your responsibility to have a readable file.

I've had over the years-- not at MIT, I must admit-- but I've had students who actually write the doc file and think that converting it to a PDF consists in erasing the extension doc and replacing it by PDF. This is not how it works. So I'm sure you will figure it out. Please try to keep them letter-sized. This is not a strict requirement, but I don't want to see thumbnails, either.

You are allowed to have two late homeworks. And by late, I mean 24 hours late. No questions asked. You submit them, this will be counted. You don't have to send an email to warn us or anything like this.

Beyond that, even that you have one slack for one 0 grade and slack for two late homeworks, you're going to have to come up with a very good explanation why you need actually more extensions than that, if you ever do. And particularly, you're going to have to keep track about why you've used your three options before.

There's going to be two midterms. One is October 3, and one is November 7. They're both going to be in class for the duration of the lecture.

When I say they last for an hour and 20 minutes, it does not mean that if you arrive 10 minutes before the end of lecture, you still get an hour and 20 minutes. It will end at the end of lecture time.

For this as well, no pressure. Only the best of the two will be kept. And this grade will count for 30% of the grade.

This will be closed-books and closed-notes. The purpose is for you to-- yes?

AUDIENCE: How many midterms did you say there are?

PHILIPPE RIGOLLET: Two.

AUDIENCE: You said the best of the two will be kept?

PHILIPPE RIGOLLET: I said the best of the two will be kept, yes.

AUDIENCE: So both the midterms will be kept?

PHILIPPE RIGOLLET: The best of the two, not the best two.

AUDIENCE: Oh.

PHILIPPE RIGOLLET: We will add them, multiply the number by 9, and that will be grade. No. I am trying to be nice, there's just a limit to what I can do.

All right, so the goal is for you to learn things and to be familiar with them. In the final, you will be allowed to have your notes with you. But the midterms are also a way for you to develop some mechanism so that you don't actually waste too much time on things that you should be able to do without thinking too much.

You will be allowed to cheat sheet, because, well, you can always forget something. And it will be two-sided letters sheet, and you can practice yourself as writing as small as you want. And you can put whatever you want on this cheat sheet.

All right, the final will be decided by the register. It's going to be three hours, and it's going to count for 40%. You cannot bring books, but you can bring your notes. Yes.

AUDIENCE: I noticed that the midterm dates aren't dated in the syllabus. So I wanted to make sure you know.

PHILIPPE RIGOLLET: They are not?

AUDIENCE: Yeah--

PHILIPPE RIGOLLET: Oh, yeah, there's a "1" that's missing on both of them, isn't there? Yeah, let's figure that out. The syllabus is the true one.

The slides are so that we can discuss, but the ones that's on the syllabus are the ones that count. And I think they're also posted on the calendar on Stellar as well. Any other question?

OK, so the pre-reqs here-- and who has looked at the first problem set already? OK, so those hands that are raised realize that there is a true prerequisite of probability for this class. It can be at the level of 18.600 or 604.1. I should say "B" now. It's two classes.

I will require you to know some calculus and have some notions of linear algebra, such as, what is a matrix, what is a vector, how do you multiply those things together, some notion of what orthonormal vectors are. We'll talk about eigenvectors and eigenvalues, but I remind you all of that. So this is not this strict pre-req. But if you've taken it, for example, it doesn't hurt to go back to your notes when we get closer to this chapter on principle-component analysis. The chapters, as they're listed in the syllabus, are in order, so you will see when it actually comes.

There's no required textbook. And I know you tend to not like that. You like to have your textbook to know where you're going and what we're doing.

I'm sorry, it's just this class. Either I would have to go to a mathematical statistics textbook, which is just too much, or to go to a more engineering-type statistics class, which is just too little. So hopefully, the problems will be enough for you to practice the recitations.

We'll have some problems to solve as well. And the material will be posted on the slides. So you should have everything you need. There's plenty of resources online if you want to expand on a particular topic or read it as said by somebody else.

The book that I recommend in the syllabus is this book called *All of Statistics* by Wasserman. Mainly because of the title, I'm guessing it has all of it in it. It's pretty broad. There's actually not that many.

It's more of an intro-grad level. But it's not very deep, but you see a lot of the overview. Certainly, what we're going to cover will be a subset of what's in there. The slides will be posted on Stellar before lectures before we start a new chapter and after we're done with the chapter, with the annotations, and also, with the typos corrected, like for the exam.

There will be some video lectures. Again, the first one will be posted on OCW from last year. But all of them will be available on Stellar-- of course, module technical problems.

But this is an automated system. And hopefully, it will work out well for us. So if you somehow have to miss a lecture, you can always catch it up by watching it. You can also play at that speed 0.75 in case I end up speaking too fast, but I think I've managed myself so far-- so just last warning.

All right, why should you study statistics? Well, if you read the news, you will see a lot of statistics. I mentioned machine learning. It's built on a lot of statistics.

If I were to teach this class 10 years ago, I would have to explain to you that data collection and making decisions based on data was something that made sense. But now, it's almost in our life. We're used to this idea that data helps in making decisions.

And people use data to conduct studies. So here, I found a bunch of press titles that-- I think the key word I was looking for was "study finds"-- if I want to do this. So I actually did not bother doing it again this year. This is all 2016, 2016, 2016.

But the key word that I look for is usually "study find"-- so a new study find-- traffic is bad for your health. So we had to wait for 2016 for data to tell us that. And there's a bunch of other slightly more interesting ones. For example, one that you might find interesting is that this study finds that students benefit from waiting to declare a major.

Now, there's a bunch of press titles. There one in the MIT News that finds brain connections, key to reading. And so here, we have an idea of what happened there.

Some data was collected. Some scientific hypothesis was formulated. And then the data was here to try to prove or disprove this scientific hypothesis. That's the usual scientific process.

And we need to understand how the scientific process goes, because some of those things might be actually questionable. Who is 100% sure that study finds that students-- do you think that you benefit from waiting to declare a major? Right I would be skeptical about this. I would be like, I don't want to wait to declare a major.

So what kind of thing can we bring? Well maybe this study studied people that were different from me. Or maybe the study finds that this is beneficial for a majority of people. I'm not a majority. I'm just one person.

There's a bunch of things that we need to understand what those things actually mean. And we'll see that those are actually not statements about individuals. They're not even statements about the cohort of people they've actually looked at. They're statements about a parameter of a distribution that was used to model the benefit of waiting.

So there's a lot of questions. And there are a lot of layers that come into this. And we're going to want to understand what was going on in there and try to peel it off and understand what assumptions have been put in there.

Even though it looks like a totally legit study, out of those studies, statistically, I think there's going to be one that's going to be wrong. Well, maybe not one. But if I put a long list of those, there would be a few that would actually be wrong. If I put 20, there would definitely be one that's wrong.

So you have to see that. Every time you see 20 studies, one is probably wrong. When there are studies about drug effects, out of a list of 100, one would be wrong. So we'll see what that means and what I mean by that. Of course, not only studies that make discoveries are actually making the press titles. There's also the press that talks about things that make no sense.

I love this first experiment-- the salmon experiment. Actually, it was a grad student who came to a neuroscience poster session, pulled out this poster, and explained the scientific experiment that he was conducting, which consisted in taking a previously frozen and thawed salmon, putting it in an MRI, showing it pictures of violent images, and recording its brain activity. And he was able to discover a few voxels that were activated by those violent images. And can somebody tell me what happened here? Was the salmon responding to the violent activity?

Basically, this is just a statistical fluke. That's just randomness at play. There's so many voxels that are recorded, and there's so many fluctuations. There's always a little bit of noise when you're in those things, that some of them, just by chance, got lit up. And so we need to understand how to correct for that.

In this particular instance, we need to have tools that tell us that, well, finding three voxels that are activated for that many voxels that you can find in the salmon's brain is just too small of a number. Maybe we need to find a clump of 20 of them, for example. All right, so we're going to have mathematical tools that help us find those particular numbers.

I don't know if you ever saw this one by John Oliver about phacking. Or actually, it said p-hacking. Basically, what John Oliver is saying is actually a full-length-- like there's long segments on this. And he was explaining how there's a sociology question here about how there's a huge incentive for scientists to publish results. You're not going to say, you know what? This year, I found nothing.

And so people are trying to find things. And just by searching, it's as if they were searching for all the voxels in a brain until they find one that was just lit up by chance. And so they just run all these studies. And at some point, one will be right just out of chance.

And so we have to be very careful about doing this. There's much more complicated problems associated to what's called p-hacking, which consists of violating the basic assumptions, in particular, looking at the data, and then formulating your scientific assumption based on data, and then going back to it. Your idea doesn't work. Let's just formulate another one. And if you are doing this, all bets are off.

The theory that we're going to develop is actually for a very clean use of data, which might be a little unpleasant. If you've had an army of graduate students collecting genomic data for a year, for example, maybe you don't want to say, well, I had one hypothesis that didn't work. Let's throw all the data into the trash. And so we need to find ways to be able to do this.

And there's actually a course been taught at BU. It's still in its early stages, but something called "adaptive data analysis" that will allow you to do these kind of things. Questions?

OK, so of course, statistics is not just for you to be able to read the press. Statistics will probably be used in whatever career path you choose for yourself. It started in the 10th century in Netherlands for hydrology.

Netherlands is basically under water, under sea level. And so they wanted to build some dikes. But once you're going to build a dike, you want to make sure that it's going to sustain some tides and some floods.

And so in particular, they wanted to build dikes that were high enough, but not too high. You could always say, well, I'm going to build a 500-meter dike, and then I'm going to be safe. You want something that's based on data. You want to make sure.

And so in particular, what did they do? Well, they collected data for previous floods. And then they just found a dike that was going to cover all these things.

Now, if you look at the data they probably had, maybe it was scarce. Maybe they had 10 data points. And so for those data points, then maybe they wanted to sort of interpolate between those points, maybe extrapolate for the larger one. Based on what they've seen, maybe they have chances of seeing something which is even larger than everything they've seen before. And that's exactly the goal of statistical modeling-- being able to extrapolate beyond the data that you have, guessing what you have not seen yet might happen.

When you buy insurance for your car, or your apartment, or your phone, there is a premium that you have to pay. And this premium has been determined based on how much you are, in expectation, going to cost the insurance. It says, OK, this person has, day a 10% chance of breaking their iPhone. An iPhone costs that much to repair, so I'm going to charge them that much. And then I'm going to add an extra dollar for my time.

That's basically how those things are determined. And so this is using statistics. This is basically where statistics is probably mostly used. I was personally trained as an actuary. And that's me being a statistician at an insurance company.

Clinical trials-- this is also one of the earliest success stories of statistics. It's actually now widespread. Every time a new drug is approved for market by the FDA, it requires a very strict regimen of testing with data, and control group, and treatment group, and how many people you need in there, and what kind of significance you need for those things. In particular, those things look like this, so now it's 5,000 patients.

It depends on what kind of drug it is, but for, say, 100 patients, 56 were cured, and 44 showed no improvement. Does the FDA consider that this is a good number? Do they have a table for how many patients were cured? Is there a placebo effect? Do I need a control group of people that are actually getting a placebo?

It's not clear, all these things. And so there's a lot of things to put into place. And there's a lot of floating parameters. So hopefully, we're going to be able to use statistical modeling to shrink it down to a small number of parameters to be able to ask very simple questions.

"Is a drug effective" is not a mathematical equation. But "Is p larger than 0.5?" is a mathematical question And that's essentially we're going to be doing. We're going to take this, is a drug effective, to reducing to, is a variable larger than 0.5?

Now, of course genetics are using that. That's typically actually the same size of data that you would see for FMRI data. So this is actually a study that I found.

You have about 4,000 cases of Alzheimer's and 8,000 control. So people without Alzheimer's-- that's what's called a control. That's something just to make sure that you can see the difference with people that are not affected by either a drug or a disease.

Is the gene APOE associated with Alzheimer's disease? Everybody can see why this would be an important question. We now have it crisper. It's targeted to very specific genes.

If we could edit it, or knock it down, or knock it up, or boost it, maybe we could actually have an impact on that. So those are very important questions, because we have the technology to target those things. But we need the answers about what those things are.

And there's a bunch of other questions. The minute you're going to talk to biologists about say, I can do that. They're going to say, OK, are there any other genes within the genes, or any particular snips that I can actually look at? And they're looking at very different questions.

And when you start asking all these questions, you have to be careful, because you're reusing your data again. And it might lead you to wrong conclusions. And those are all over the place, those things. And that's why they go all the way to John Oliver talking about them.

Any questions about those examples? So this is really a motivation. Again, we're not going to just take this data set of those cases and look at them in detail.

So what is common to all these examples? Like, why do we have to use statistics for all those things? Well, there's the randomness of the data.

There's some effect that we just don't understand-- for example, the randomness associated with the lining up of some voxels. Or the fact that as far as the insurance is concerned whether you're going to break your iPhone or not is essentially a coin toss. Fully, it's biased. But it's a coin toss.

From the perspective of the statistician, those things are actually random events. And we need to tame this randomness, to understand this randomness. Is this going to be a lot of randomness? Or is it going to be a little randomness?

Is it going to be something that's like, out of their people-- let's see, for example, for the floods. Were the floods that I saw consistently almost the same size? It was almost a rounding error, or they're just really widespread. All these things, we need to understand so we can understand how to build those dikes or how to make decisions based on those data. And we need to understand this randomness.

OK, so the associated questions to randomness were actually hidden in the text. So we talked about the notion of average. Right, so as far as the insurance is concerned, they want to know in average with the probability is. Like, what is your chance of actually breaking your iPhone? And that's what came in this notion of fair premium.

There's this notion of quantifying chance. We don't want to talk maybe only about average, maybe you want to cover say 99% percent of the floods. So we need to know what is the height of a flood that's higher than 99% of the floods. But maybe there's 1% of them, you know. When doomsday comes, doomsday comes. Right, we're not going to pay for it. All right, so that's most of the floods.

And then there's questions of significance, right? So you know I give this example, a second ago about clinical trials. I give you some numbers. Clearly the drug cured more people than it did not. But does it mean that it's significantly good, or was this just by chance. Maybe it's just that these people just recovered. It's like you know curing a common cold. And you feel like, oh I got cured. But it's really you waited five days and then you got cured.

All right, so there's this notion of significance, of variability. All these things are actually notions that describe randomness and quantify randomness into simple things. Randomness is a very complicated beast. But we can summarize it into things that we understand. Just like I am a complicated object. I'm made of molecules, and made of genes, and made of very complicated things. But I can be summarized as my name, my email address, my height and my weight, and maybe for most of you, this is basically enough. You will recognize me without having to do a biopsy on me every time you see me.

All right, so, to understand randomness you have to go through probability. Probability is the study of randomness. That's what it is. That's what the first sentence that a lecturer in probability will say. And so that's why I need the pre-requisite, because this is what we're going to use to describe the randomness. We'll see in a second how it interacts with statistics.

So sometimes, and actually probably most of the time throughout your semester on probability, randomness was very well understood. When you saw a probability problem, here was the chance of this happening, here was the chance of that happening. Maybe you had more complicated questions that you had some basic elements to answer.

For example, the probability that I have HBO is this much. And the probability that I watch *Game of Thrones* is that much. And given that I play basketball what is the probability-- you had all these crazy questions, but you were able to build them. But all the basic numbers were given to you. Statistics will be about finding those basic numbers.

All right so some examples that you've probably seen were dice, cards, roulette, flipping coins. All of these things are things that you've seen in a probability class. And the reason is because it's very easy to describe the probability of each outcome. For a die we know that each face is going to come with probably 1/6. Now I'm not going to go into a debate of whether this is pure randomness or this is determinism. I think as a model for actual randomness a die is a pretty good number, flipping a coin is a pretty good model. So those are actually a good thing.

So the questions that you would see, for example, in probabilities are the following. I roll one die. Alice gets $1 if the number of dots is less than three. Bob gets $2 if the number of dots is less than two. Do you want to be Alice or Bob given that your role is actually to make money.

Yeah, you want to be Bob, right? So let's see why. So if you look at the expectation of what Alice makes. So let's call it a. This is $1, with probability 1/2. So 3/6, that's 1/2. And the expectation of what Bob makes, this is $2 with probably 2/6 and that's 2/3. Which is definitely larger than 1/2. So Bob's expectations actually a bit higher.

So those are the kind of questions that you may ask with probability. I described to you exactly, you use the fact that the die would get less than three dots, with probability one half. We knew that. And I didn't have to describe to you what was going on there. You didn't have to collect data about a die. Same thing, you roll two dice. You choose a number between 2 and 12 and you win $100 if you choose the sum of the two dice. Which number do you pick? What?

AUDIENCE: 7.

PHILIPPE RIGOLLET: 7. Why 7?

AUDIENCE: It's the most likely.

PHILIPPE RIGOLLET: That's the most likely one, right? So your gain here will be $100 times the probability that the sum of the two dice, let's say x plus y, is equal to your little z where a little z is the number you pick. So 7 is the most likely to happen and that's the one that maximizes this function of z. And for this you need to study a more complicated function. But it's a function that enables two die. But you can compute the probability that x plus y is equal to z, for every z between 2 and 12. So you know exactly what the probabilities are and that's how you start probability.

So here that's exactly what I said. You have a very simple process that describes basic events. Probability 1/6 for each of them. And then you can build up on that, and understand probably of more complicated events. You can throw some money in there. You can be build functions. You can do very complicated things building on that.

Now if I was a statistician, a statistician would be the guy who just arrived on earth, had never seen a die and needs to understand that a die come up with probably 1/6 on each side. And the way he would do it is just to roll the die until he get some counts and tries to estimate those. And maybe that guy would come and say, well, you know, actually, the probability that I get a 1 is 1/6 plus 0.001 and the probability that I get a 2 is 1/6 minus 0.005. And there would be some fluctuations around this.

And it's going to be his role as a statistician to say, listen, this is too complicated of a model for this thing. And these should all be the same numbers. Just looking at data, they should be all the same numbers. And that's part of the modeling. You make some simplifying assumptions that essentially make your questions more accurate.

Now, of course, if your model is wrong, if it's not true that all the faces arrive with the same probability, then you have a model error here. So we will be making model errors. But that's going to be the price to pay to be able to extract anything from our data.

So for more complicated processes, so of course nobody's going to waste their time rolling dice. I mean, I'm sure you might have done this in AP stat or something. But the need is to estimate parameters from data.

All right, so for more complicated things you might want to estimate some density parameter on a particular set of material. And for this maybe you need to beam something to it, and measure how fast it's coming back. And you're going to have some measurement errors. And maybe you need to do that several times and you have a model for the physical process that's actually going on. And physics is usually a very good way to get models for engineering perspective.

But there's models for sociology where we have no physical system, right. God knows how people interact. And maybe I'm going to say that the way I make friends is by first flipping a coin in my pocket. And with probability 2/3, I'm going to make my friend at work. And with probability 1/3 I'm going to make my friend at soccer.

And once I make my friends at soccer-- I decide to make my friend soccer. Then I will face someone who's flipping the same coin with maybe be slightly different parameters. But those things actually exist. There's models about how friendships are formed. And the one I described is called the mixed-membership model. So those are models that are sort of hypothesized. And they're more reasonable than taking into account all the things that made you meet that person at that particular time.

So the goal here-- so based on data now, once we have the model is going to be reduced to maybe two, three, four parameters, depending on how complex the model is. And then your goal will be to estimate those parameters.

So sometimes the randomness we have here is real. So there's some true randomness in some surveys. If I pick a random student, as long as I believe that my random number generator that will pick your random ID is actually random, there is something random about you. The student that I pick at random will be a random student. The person that I call on the phone is a random person. So there's some randomness that I can build into my system by drawing something from a random number generator.

A biased coin is a random thing. It's not a very interesting random thing. But it is a random thing. Again, if I wash out the fact that it actually is a deterministic mechanism. But at a certain accuracy, a certain granularity, this can be thought of as a truly random experiment.

Measurement error for example, if you by some measurement device. or some optics device, for example. You will have like standard deviation and things that come on the side of the box. And it tells you, this will be making some measurement error. And it's usually thermal noise maybe, or things like this. And those are very accurately described by some random phenomenon.

But sometimes, and I'd say most times, there's no randomness. There's no randomness. It's not like you breaking your iPhone is a random event. This is just something that we sweep-- randomness is a big rug under which we sweep everything we don't understand. And we just hope that in average we've captured, the average effect of what's going on. And the rest of it might fluctuate to the right, might fluctuate to the left. But what remains is just sort of randomness that can be averaged out.

So, of course, this is where the leap of faith is. We do not know whether we were correct of doing this. Maybe we make some huge systematic biases by doing this. Maybe we forget a very important component. Right, for example, if I have-- I don't know, let's think of something-- a drug for breast cancer.

All right, and I throw out the fact that my patient is either a man or woman. I'm going to have some serious model biases. Right. So if I say I'm going to collect a random and patient. And said I'm going to start doing this. There's some information that I really need, clearly, to build into my model.

And so the model should be complicated enough, but not too complicated. Right so it should take into account things there will systematically be important.

So, in particular, the simple rule of thumb is, when you have a complicated process, you can think of it as being a simple process and some random noise. Now, again, the random noise is everything you don't understand about the complicated process. And the simple process is everything you actually do.

So good modeling, and this is not where we'll be seeing in this class, consistent choosing plausible simple models. And this requires a tremendous amount of domain knowledge. And that's why we're not doing it in this class. This is not something where I can make a blanket statement about making good modeling.

You need to know, if I were a statistician working on a study, I would have to grill the person in front of me, the expert, for two hours to know, but how about this? How about that? How does this work? So it requires to understand a lot of things.

There's this famous statistician to whom this sentence is attributed, and it's probably not his then, but Tukey said that he loves being a statistician, because you get to play in everybody's backyard. Right, so you get to go and see people. And you get to understand, at least to a certain extent, what their problems are. Enough that you can actually build a reasonable model for what they're actually doing.

So you get to do some sociology. You get to do some biology. You get to do some engineering. And you get to do a lot of different things. Right, so he was actually at some point predicting the presidential election.

So, you see, you get to do a lot of different things. But it requires a lot of time to understand what problem you're working on. And if you have a particular application in mind you're the best person to actually understand this. So I'm just going to give you the basic tools.

So this is the circle of trust. No, this is really just a simple graphic that tells you what's going on. When you do probability, you're given the truth. Somebody tells you what die God is rolling. So you know exactly what the parameters of the problems are. And what you're trying to do is to describe what the outcomes are going to be.

You can say, if you're rolling a fair die, you're going to have 1/6 of the time in your data you're going to have one. 1/6 of the time you're going to have to have two. And so you can describe-- if I told you what the truth is, you could actually go into a computer, either generate some data. Or you could describe to me some more macro properties of what the data would be like.

Oh, I would see a bunch of numbers that would be centered around 35, if I drew from a Gaussian distribution centered at 35. Right, you would know this kind of thing. I would know that it's very unlikely that if my Gaussian has standard deviation-- is centered on 0, say, with standard deviation 3. It's very unlikely that I will see numbers below minus 10 in above 10, right? You know this, that you basically will not see them.

So you know from the truth, from the distribution of a random variable that does not have mu or sigmas, really numbers there. You know what data, you're going to be having. Statistics is about going backwards. It's saying, if I have some data, what was the truth that generated it. And since there are so many possible truths, Modeling says you have to pick one of the simpler possible truths, so that you can average out.

Statistics basically means averaging. You're averaging when you do statistics. And averaging means that if I say that I received-- so if I collect all your GPAs, for example. And my model is that the possible GPAs are any possible numbers. And anybody can have any possible GPA. This is going to be a serious problem.

But if I can summarize those GPAs into two numbers, say, mean and standard deviation, than I have a pretty good description of what is going on, rather than having to have to predict the full list. Right, if I learn a full list of GPAs and I say, well this was the distribution. Then it's not going to be of any use for me to predict what the GPA would be, or some random student walking in, or something like this.

So just to finish my rant about probability versus statistics, this is a question you would see in a probability-- this is a probabilistic question, and this is a statistical question. The probabilistic question is, previous studies showed that the drug was 80% effective. So you know that. This is the effectiveness of the drug. It's given to you. This is how your problem starts. Then we can anticipate that, for a study on 100 patients, in average, 80 be cured. And at least 65 will be cured with 99% chances.

So again these are not-- I'm not predicting on 100 patients exactly the number of them they're going to be cured. And the number of them that are not. But I'm actually sort of predicting what things are going to look like on average, or some macro properties of what my data sets will look like.

So with 99 percent chances, that means that in 99.99% of the data sets you will draw from this particular draw. 99.99% of the cohort of 100 patients to whom you administer this drug, I will be able to conclude that at least 65 of them will be cured, on 99.99% percent of those data sets.

So that's a pretty accurate prediction of what's going to happen. Statistics is the opposite. It says, well, I just know that 78 out of 100 were cured. I have only one data set. I cannot make predictions for all data sets. But I can go back to the probability, make some inference about what my probability will look like, and then say, OK, then I can make those predictions later on.

So when I start with 78/100 then maybe I'm actually, in this case, I just don't know. My best guess here is that I'm confident I have to add the extra error that I bet you making by predicting that here, the drug is not 80% effective but 78% effective. And they need some error bars around this, that will hopefully contain 80%, and then based on those error bars I'm going to make slightly less precise predictions for the future.

So, to conclude, so this was, why statistics? So what is this course about? It's about understanding the mathematics behind statistical methods. It's more of a tool. We're not going to have fun and talk about algebraic geometry just for fun in the middle of it. So it justifies quantitative statements given some modeling assumptions, that we will, in this class, mostly admit that the modeling assumptions are correct.

| the first part-- in this introduction, we will go through them because it's very easy to forget what the assumptions are actually making. But this will be a pretty standard thing. The words you will hear a lot are IID-- independent and identically distributed-- that means that your data is basically all the sams. And one data point is not impacting another data point.

Hopefully we can describe some interesting mathematics arising in statistics. You know, if you've taken linear algebra, maybe we can explain to you why. If you've done some calculus, maybe we can do some interesting calculus. We'll see how in the spirit of applied math those things answer interesting questions.

And basically we'll try to carve out a math toolbox that's useful for us statistics. And maybe you can extend it to more sophisticated methods that we did not cover in this class. In particular in the immersion learning class, hopefully you'll be able to have some statistical intuition about what is going on.

So what this course is not about, it's not about spending a lot of time looking at data sets, and trying to understand some statistical thinking kind of questions. So this is more of an applied statistical perspective on things, or more modeling. So I'm going to typically give you the model. And say this is a model. And this is how we're going to build an estimator in the framework of this model.

So for example, 18.075, to a certain extent, is called "Statistical Thinking and Data Analysis." So I'm hoping there is some statistical thinking in there. We will not talk about software implementation. Unfortunately, there's just too little time in a semester. There's other courses that are giving you some overview. So the main software these days are R is the leading software I'd say in statistics, both in academia and industry, lots of packages, one every day that's probably coming out.

But there's other things, right, so now Python is probably catching up with all these scikit-learn packages that are coming up. Julia has some statistics in there, but it really if you were to learn a statistical software, let's say you love doing this, this would be the one that would prove most useful for you in the future. It does not scale super well to high dimensional data.

So there is a class an IDSS that actually uses R. It's called IDS 0.12, I think it's called "Statistics, Computation, and Applications," or something like this. I'm also preparing, with Peter Kempthorne, a course called "Computational Statistics." It's going to be offered this Spring as a special topics. And so Peter Kempthorne will be teaching it. And this class will actually focus on using R. And even beyond that, it's not just going to be about using. It's going to be about understanding-- just the same way we we're going to see how math helps you do statistics, it's going to help see how math helps you do algorithims for statistics.

All right, so we'll talk about maximum likelihood estimator. Will need to maximize some function. There's an optimization toolbox to do that. And we'll see how we can have specialized for statistics for that, and what are the principles behind it. And you know, of course, if you've taken AP stats you probably think that stats is boring to death because it was just a long laundry-list that spent a lot of time on t-test. I'm pretty sure we're not going to talk about t-test, well, maybe once. But this is not a matter of saying you're going to do this. And this is a slight variant of it. We're going to really try to understand what's going on.

So, admittedly, you have not chosen the simplest way to get an A in statistics on campus. All right, this is not the easiest class. It might be challenging at times, but I can promise you that you will maybe suffer. But you will learn something by the time you're out of this class. This will not be a waste of your time. And you will be able to understand, and not having to remember by heart how those things actually work.

Are there any questions?

Anybody want to go to other stats class on campus? Maybe it's not too late. OK.

So let's do some statistics. So I see the time now and it's 11:56, so we have another 30 minutes. I will typically give you a three, four minute break if you want to stretch, if you want to run to the bathroom, if you want to check your texts or Instagram. There was very little content in this class, hopefully it was entertaining enough that you don't need the break. But just in the future, so you know you will have a break.

So statistics, this is how it starts, I'm French, what can I say I need to put some French words. So this is not how office hours are going to go down.

Anybody know this sculpture by a Rodin, *The Kiss.* Maybe probably *The Thinker* is more famous. But this is actually a pretty famous one. But is it really this one, or is it this one.

Anybody knows which one it is?

This one? Or this one?

AUDIENCE: The previous.

PHILIPPE RIGOLLET: What's that?

AUDIENCE: This one.

PHILIPPE RIGOLLET: It's this one.

AUDIENCE: Final answer.

PHILIPPE RIGOLLET: Yeah, who votes for this one. OK. Who votes for that one? Thank you. I love that you do not want to pronounce yourself with no data actually to make any decision. This is a total coin toss right. Turns out that there is data, and there is in the very serious journal *Nature,* someone published a very serious paper which actually looks pretty serious.

If you look at it, it's like "Human Behavior: Adult persistence of head-turning symmetry," is a lot of fancy words in there. And this, I'm not kidding you, this study is about collecting data of people kissing, and knowing if they bend their head to the right or if they bend they head to the left. And that's all it is. And so a neonatal right-side preference makes a surprising romantic reappearance in later life. There's an explanation for it.

All right, so if we follow this *Nature* which one is the one.

This one? Or this one?

This one, right? Head to the right. And to be fair, for this class I was like, oh, I'm going to go and show them what Google Images does. When you Google kissing couple, it's inappropriate after maybe the first picture. And so I cannot show you this. But you know you can check for yourself.

Though I would argue, so this person here actually went out in airports and took pictures of strangers kissing and collecting data. And can somebody guess why did he just not stay home and collect data from Google Images by just googling kissing couples. What's wrong with this data? I didn't know actually before I actually went on Google Images.

AUDIENCE: It can be altered?

PHILIPPE RIGOLLET: What was that?

AUDIENCE: It can be altered.

PHILIPPE RIGOLLET: It can be altered. But, you know, who would want to do this? I mean there's no particular reason why you would want to flip an image before putting it out there. I mean, you might, but you know maybe they want to hide the brand of your Gap shirt or something.

AUDIENCE: I guess the people who post pictures of themselves kissing on Google Images are not representative of the general population.

PHILIPPE RIGOLLET: Yeah, that's very true. And actually it's even worse than that. The people who post pictures of themselves, are not posting pictures of themselves or putting pictures of the people that they took a picture of. And there usually is a stock watermark on this. And it's basically stock images. Those are actors, and so they've been directed to kiss and this is not a natural thing to do. And actually, if you go to Google Images-- and I encourage you to do this, unless you don't want to see inappropriate pictures, and they're mightily inappropriate.

And basically you will see that this study is actually not working at all. I mean, I looked briefly. I didn't actually collect numbers. But I didn't find a particular tendency to bend right. If anything, it was actually probably the opposite. And it's because those people were directed to do it. They just don't actually think about doing it.

And also because I think you need to justify writing in your paper more than, I sat in front of my computer. So again, this first sentence here, a neonatal right-side preference-- "is there a right side preference?" is not a mathematical question. But we can start saying, let's blah, and put some variables, and ask questions about those variables. So you know x is actually not a variable that's used very much in statistics for parameters. But p is one, for parameter.

And so you're going to take your parameter of interest, p, As here is going to be the proportion of couples. And that's among all couples. So here, if you talk about statistical thinking, there would be a question about what population this would actually be representative of.

| usually this is a call to your-- sorry, I should not forget this word it's important for you. OK, I forget this word. So this is-- OK,

So if you look at this proportion, maybe these couples that are in the study might be representative only of couples in airports. Maybe they actually put on a show for the other passengers. Who knows? You know, like, oh, let's just do it as well. And just like the people in Google Images they are actually doing it. So maybe you want to just restrict it. But of course clearly if it's appearing in *Nature,* it should not be only about couples in airports. It's supposedly representative of all couples in the world.

And so here let's just keep it vague, but you need to keep in mind what population this is actually making a statement about. So you have this full population of people in the world. Right, so those are all the couples. And this person went ahead and collected data about a bunch of them.

And we know that, in this thing, there's basically a proportion of them, that's like p, and that's the proportion of them that's bending their head to the right. And so everybody on this side is bending their heads right. And hopefully we can actually sample this thing you're informing. That's basically the process that's going on.

So this is the statistical experiment. We're going to observe n kissing couples. So here we're going to put as many variables as we can. So we don't have to stick with numbers. And then we'll just plug in the numbers. n kissing couples, and n is also, in statistics, by the way, n is the size of your sample 99.9% of the time. And collect the value of each outcome.

So we want numbers. We don't want right or left. So we're going to code them by 0 and 1, pretty naturally. And then we're going to estimate p which is unknown. So p is this area. And we're going to estimate it simply by the proportion of right So the proportion of crosses that actually fell in the right side.

So in this study what you will find is that the numbers that were collected were 124 couples, and that, out of those 124, 80 of them turned their head to the right. So, p hat is a proportion. How do we do it? Well, you don't need statistics for that. You're going to see 80 divided by 124. And you will find that in this particular study 64.5% of the couples were bending their heads to the right. That's a pretty large number, right?

The question is if I picked another 124 couples, maybe at different airports, different times, would I see same number? Would this number be all over the place? Would it be sometimes very close to 120, or sometimes for close to 10? Or would it be-- is this number actually fluctuating a lot.

And so, hopefully not too much, 64.5 percent is definitely much larger than 50%. And so there seems to be this preference. Now we're going to have to quantify how much of this preference. Is this number significantly larger than 50%? So if our data, for example, was just three couples. I'm just going there, I'm going to Logan. I call it, I do right, left right.

And then I see-- see what's the name of the fish place there? I go to I go to Wahlburgers at Logan and I'm like, OK, I'm done for the day. I collect this data. I go home, and I'm like, wow, 66.7% to the right. That's a pretty big number. It's even farther from 50% than this other guy. So I'm doing even better.

But of course you know that this is not true. Three people is definitely not representative. If I stopped at the first one, I would have actually-- at the first two, I would have even 100%.

So the question that statistics is going to help us answer is, how large should the sample be? For some reason, I don't know if you guys receive this, I'm an affiliate with the Broad Institute, and since then I receive one email per day that says, sample size determination-- how large should your sample be? Like, I know how large should with my sample be. I've taken 18.650 multiple times.

And so I know, but the question is-- is 124 a large enough number or not? Well, the answer is actually, as usual, it depends. It will depend on the true unknown value of p. But from those particular values that we got, so 120 and-- how many couples was there? 80? We actually can make some question.

So here we said that 80 was larger than 50-- was allowing us to conclude at 64.5%. So it could be one reason to say that it was larger than 50%. 50% of 124 is 62.

So the question is, would I be would I be willing to make this conclusion at 63? Is that a number that would convince you? Who would be convinced by 63? who would be convinced by 72? Who would be convinced by 75? Hopefully the number of hands that are raised should grow. Who would be convinced by 80?

All right, so basically those numbers actually don't come from anywhere. This 72 would be the number that you would need for a study-- most statistical studies would be the number that they would retain. That's not for 124. You would need to see 72 that turn their head right to actually make this conclusion. And then 75--

So we'll see that there's many ways to come to this conclusion because, as you can see, this was published in *Nature* with 80. So that was OK. So 80 is actually a very large number. This is 99 point-- this 99% -- no, so this is 95% confidence.

This is 99% confidence. And this is 99.9% percent confidence. So if you said 80 you're a very conservative person. Starting at 72, you can start making this conclusion.

To understand this, we need to do our little mathematical kitchen here, and we need to do some modeling. So we need to understand by modeling-- we need understand what random process we think this data is generating from. So it's going to have some unknown parameters, unlike in probability. But we need to have just basically everything written except for the values of the parameters.

When I said a die is coming uniformly with probably 1/6 then I need to have, say maybe with probability-- maybe I should say here are six numbers, and I need to just fill those numbers.

So for i equal 1 to n, I'm going to define Ri to be the indicator. An indicator is just something that takes value 1 if something is true, and 0 if not. So it's an indicator that i-th couple turns the head to the right. So, Ri, so it's indexed by i. And it's one if the i-th couple turns their head to the right, and 0 if it's-- well actually, I guess they can probably kiss straight, right? So that would be weird, but they might be able to do this. So let's say not right.

Then the estimator of p, we said, was p hat. It was just the ratio of two numbers. But really what it is is I count, I sum those Ri's. Since I only add those that take value 1, what this is is-- this sum here is actually just counting the number of 1's. Which is another way to say it's counting the number of couples that are kissing to the right.

And here I don't even have to tell you anything about the numbers or anything. I can only keep track of-- first couple is a 0 second couple is a 1, third couple is 0. The data set-- you can actually find it online-- is actually a sequence of 0's and 1's. Now clearly for the question that we're asking about this proportion, I don't need to keep track of all this information. All I need to keep track of is the number of 0's and the number of 1's. Those are completely interchangeable. There's no time effect in this. The first couple is no different than the 15th couple.

So we call this Rn bar. That's going to be a very standard notation that we use. R might be replaced by other letters like x-- so xn bar, yn bar. And this thing essentially means that I average the R's, or the Ri's over n of them. And the bar means the average. So I divide by n the total number of 1's. So here this sum was equal to 80 in our example and n was equal to 124.

Now this is an estimator. So an estimator is different from an estimate. An estimate is a number. My estimate was 64.5. My estimator is this thing where I keep all the variables free. And in particular, I keep those variables to be random because I'm going to think of a random couple kissing left to right as the outcome of a random process, just like flipping a coin be getting heads or tails.

And so this thing here is a random variable, Ri. And this average is, of course, an average of random variables. It's itself a random variable. So an estimator is a random variable. An estimate is the realization of a random variable, or, in other words, is the value that you get for this random variable once you plug in the numbers that you've collected.

So I can talk about the accuracy of an estimator. Accuracy means what? Well, what would we want for an estimator? Maybe we won't want it to fluctuate too much. It's a random variable. So I'm talking about the accuracy of a random variable. So maybe I don't want it to be too volatile.

I could have one estimator which would be-- just throw out 182 couples, keep only 2 and average those two numbers. That's definitely a worse estimator than keeping all of the 124. So I need to find a way to say that. And what I'm going to be able to say is that the number is going to be fluctuating. If I take another two couples, I'm going to be I'm probably going to get a completely different number. But if I take another 124 couples two days later, maybe I'm going to have a very number that's very close to 64.5%.

So that's one way. The other thing we would like about this estimator it's actually-- maybe it's not too volatile-- but also we want it to be close to the number that we're looking for. Here is an estimator. It's a beautiful variable. 72%, that's an estimator. Go out there just do your favorite study about drug performance. And then they're going to call you, MIT student taking statistics, they say, so how are you going to build your estimator? We've collected those 5,000 or something like that.

I'm just going to spit out 72%. Whatever the data says, that's an estimator. It's a stupid estimator but it is an estimator. But this is estimator is very not volatile. Every time you're going to have a new study, even if you change fields, it's still going to be 72%. This is beautiful. And the problem is that's probably not very close to the value you're actually trying to estimate.

So we need two things. We need are estimated to be a random variable. So think in terms of densities. We want the density to be pretty narrow. We want this thing to have very little-- so this is definitely better than this. But also, we want the number that we're interested in, p, to be very close to this-- to be close to the values that this thing is likely to take. If p is here, this is not very good for us.

So that's basically the things we're going to be looking at. The first one is referred to as variance. The second one is referred to as bias. Those things come all over in statistics.

So we need to understand a model. So here's the model that we have for this particular problem. So we need to make assumptions on the observations that we see. So we said we're going to assume that the random variable-- that's not too much of a leap of faith. We're just sweeping under the rug everything thing we don't understand about those couples.

And the assumption that we make is that Ri is a random variable. This one you will forget very soon. The second one is that each of the Ri's is-- so it's a random variable that takes value 0 or 1. Anybody can suggest the distribution for this random variable?

AUDIENCE: Bernoulli.

PHILIPPE RIGOLLET: What?

AUDIENCE: Bernoulli.

PHILIPPE RIGOLLET: Bernoulli, right? And it's actually beautiful. This is where you have to do the least statistical modeling. A random variable that takes value 0 or 1 is always a Bernoulli. That's the simplest variable you can ever think of. Any variable that takes only two possible values can be reduced to a Bernoulli. OK, so this is a Bernoulli.

And here we make the assumption that it actually takes parameter p. And there's an assumption here. Anybody can tell me what the assumption is?

AUDIENCE: It's the same.

PHILIPPE RIGOLLET: Yeah, it's same, right? I could have said p i, but it's p. And that's where I'm going to be able to start getting to do some statistics. It's that I'm going to start to be able to pull information across all my guys. If I assume that they're all pi's completely uncoupled with each other. Then I'm in trouble. There's nothing I can actually get.

And then I'm going to assume that those guys are mutually independent. And most of the time they will just say independent. Meaning that, it's not like all these guys called each other and it's actually a flash mob. And they were like, let's all turn our left side to the left. And then this is definitely not going to give you a valid conclusion.

So, again. randomness is a way of modeling lack of information. Here there is a way to figure it out. Maybe I could have followed all those guys, and knew exactly what they were-- maybe I could have looked at pictures of them in the womb and guess how they were turning-- by the way that's one of the conclusions, they're guessing that we turn our head to the right because our head is turned to the right in the womb. So we don't know what goes on in the kissers minds. And there's, you know, physics, sociology. There's a lot of things that could help us, but it's just too complicated to keep track of, or too expensive for many instances

Now again, the nicest part of this modeling was the fact that Ri's take only two values, which mean that this conclusion that they were Bernoulli was totally free for us. Once we know it's a random variable, it's a Bernoulli. Now they could have been, as we said, they could have been a Bernoulli with parameter p i.

For each i, I could have put a different parameter, but I just don't have enough information. What would I have said? I would say, well the first couple turned to the right. p1 has to be one, that's my best guess. The second couple kiss to the left, well, p2 should be 0, that's my best guess.

And so basically I need to have to be able to average my information. And the way I do it is by coupling all these guys, pi's to be the same p for all i. OK, does it make sense? Here what I am assuming is that my population is homogeneous. Maybe it's not. Maybe I could actually look at a finer grain, but I'm basically making a statement about a population.

And so maybe you kiss to the left, and then you're not-- I'm not making a statement about a person individually, I'm making a statement about the overall population.

Now independence is probably reasonable, right? This person just went and know can seriously hope that these couples did not communicate with each other. Or that you know Tanya did not text that we should all turn our head to the left now. And there's no external stimulus that forces people to do something different.

OK, so-- sorry about that. Since we have about less than 10 minutes. Let's do a little bit of exercises, is that OK with you? So I just have some exercises so we can see what an exercise going to look like. This is sort of similar to the exercises you will see with me. We should do it together, OK?

So now we're going to have-- I have a test. So that's an exam in probability. OK. And I'm going to have 15 students in this test. And hopefully, this should be 15 grades that are representative of the grades of all a large class.

Right, so if you go you know 18.600, it's a large class, there's definitely more than 15 students. And maybe, just by sampling 15 students at random, I want to have an idea of what my grade distribution will look like. I'm grading them, I want to make an educated guess.

So I'm going to make some modeling assumptions for those guys. So here, 15 students and the grades are x1 to x15. Just like we had R1, R2, all the way to R124. Those were my Ri's. And so now I have my xi's. And I'm going to assume that xi follows a Gaussian or normal distribution with min mu and variance sigma squared.

Now this is modeling, right? Nobody told me there's no physical process that makes this happen. We know that there's something called the central limit theorem in the background that says that things tend to be Gaussian, but this is really a matter of convenience.

Actually this is, if you think about it, this is terrible because this puts non-zero probability on negative scores. I'm definitely not going to get a negative score. But you know it's good enough because they know the probabilities non-zero but it's probably 10 to the minus 12. So I would be very unlucky to see a negative score.

So here's the list of grades, so I have 65, 41, 70, 90, 58, 82, 76, 78-- maybe I should have done it with 8 --59, 59-- sitting next to each other --84, 89, 134, 51, and 72.

So those are the scores that I got. There were clearly some bonus points over there. And the question is, find estimator for mu. What is my estimator for mu? Well, an estimator, again, is something that depends on the random variable. All right, so mu is the expectation, right? So a good estimator is definitely the average score, just like we had the average of the Ri's.

Now the xi's no longer need to be 0's and 1's, so it's not going to boil down to being a number of ones divided by the total numbers. Now if I'm looking for an estimate, well, I need to actually sum those numbers and divide them by 15. So my estimate is going to be 1/15.

Then I'm going to start summing those numbers-- 65 plus 72. OK, and I can do it, and it's 67.5. This is my estimate. Now if I want to compute a standard deviation-- so let's say estimate for sigma. You've seen that before, right? An estimate for sigma is what? An estimate for sigma, we'll see methods to do this, but sigma squared is the variance, or is the expectation, of x minus expectation of x squared.

And the problem is that I don't know what those expectations are. And so I'm going to do what 99.9% percent of statistics is. And what is statistics about? What's my motto? Statistics is about replacing expectations with averages. That's what all of statistics is about. There's 300 pages in a purple book called *All of Statistics* that tells you this. All right, and then you do something fancy. Maybe you minimize something after you replace the expectation. Maybe you need to plug in other stuff. But really, every time you see an expectation, you replace it by an average.

OK let's do this. So sigma squared hat will be what? It's going to be 1 over n, sum from i equals 1 to n of xi minus-- well, here I need to replace my expectation by an average, which is really this average. I'm going to call it mu hat squared.

There, you have replaced my expectation with average. OK so the golden thing is, take your expectation and replace it with this. Frame it, get a tattoo, I don't care but that's what it is. If you remember one thing from this class, that's what it is.

Now you can be fancy, if you look at your calculator, it's going to put an n minus 1 here because it wants to be unbiased. And those are things we are going to come to. But let's say right now we stick to this. And then when I plug in my numbers. I'm going to get an estimate for sigma, which is the square root of the estimator once I plug in the numbers. And you can check that the number, you get will be 18.

So those are basic things and if you've taken any AP stats this should be completely standard to you.

Now I have another list, and I don't have time to see it. It doesn't really matter. OK, we'll do that next time. This is fine. We'll see another list of numbers and see-- we're going to think about modeling assumptions. The goal of this exercise is not to compute those things, it's really to think about modeling assumptions. Is it reasonable to think that things are IID? Is it reasonable to think that they have all the same parameters, that they're independent, et cetera,

OK so one thing that I wanted to add is, probably by tonight, so I will try to use-- in the spirit of-- I don't know what's starting to happen. In the spirit of using my iPad and fancy things, I will try to post some videos of-- for in particular, who has never used a statistical table to read, say, the quantiles of a Gaussian distribution?

OK, so there's several of you. This is a simple but boring exercise. I will just post a video on how to do this, and you will be able to find it on Stellar. It's going to take five minutes, and then you will know everything there is to know about those things but that's something you need for the first problem set.

By the way, so the problem set has 30 exercises in probability. You need to do 15. And you only need to turn in 15. You can turn in all of 30 if you want. But you need to know, by the time we hit those things, you need to know - well actually, by next week you need to know what's in there.

So if you don't have time to do all the homework, and then go back to your probability class to figure out how to do it, just do 15 easy that you can do. And return those things. But go back to your probability class and make sure that you know how to do all of them. Those are pretty basic questions, and those are things that I'm not going to slow down on. So you need to remember that the expectation of the product of independent random variables is a product of the expectations. Expectation of the sum, is the sum of the expectation. This kind of thing, which is a little silly, but it just requires you practice. So, just have fun. Those are simple exercises. You will have fun remembering your probability class.

All right, so I'll see you on Tuesday-- or Monday.